Advanced Machine Learning Summer 2019

Part 16 – Latent Variable Models III

19.06.2019

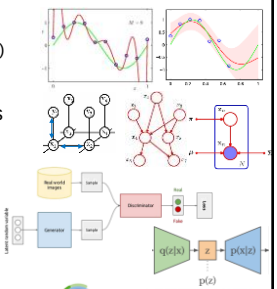
Prof. Dr. Bastian Leibe

RWTH Aachen University, Computer Vision Group
http://www.vision.rwth-aachen.de

Course Outline

- Regression Techniques
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Kernels (Kernel Ridge Regression)
- Deep Reinforcement Learning
- Probabilistic Graphical Models
  - Bayesian Networks
  - Markov Random Fields
  - Inference (exact & approximate)
  - Latent Variable Models
- Deep Generative Models
  - Generative Adversarial Networks
  - Variational Autoencoders

$f : \mathcal{X} \rightarrow \mathbb{R}$



Topics of This Lecture

- Recap: General EM
- Bayesian Estimation Revisited
  - Conjugate priors
  - Probability distributions
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
- Approximate Inference for Bayesian Mixture Models
  - Gibbs Sampler

Recap: General EM Algorithm

- Algorithm
  1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$
  2. E-step: Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
  3. M-step: Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

  where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

  4. While not converged, let $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$ and return to step 2.

Recap: The EM Algorithm in General

- Decomposition
  - Introduce a distribution $q(\mathbf{Z})$ over the latent variables. For any choice of $q(\mathbf{Z})$, the following decomposition holds

$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \parallel p)$$

  - where

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

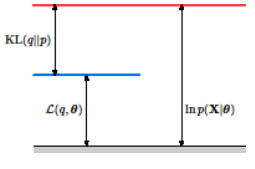
$$KL(q \parallel p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

  - $KL(q \parallel p)$ is the Kullback-Leibler divergence between the distribution $q(\mathbf{Z})$ and the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$.
  - $\mathcal{L}(q, \boldsymbol{\theta})$ is a functional of the distribution $q(\mathbf{Z})$ and a function of the parameters $\boldsymbol{\theta}$. Since KL ≥ 0, $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on $\log p(\mathbf{X}|\theta)$.

Recap: Analysis of EM

- Decomposition

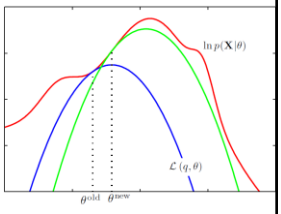
$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \parallel p)$$



- Interpretation
  - $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on $\log p(\mathbf{X}|\boldsymbol{\theta})$.
  - The approximation comes from the fact that we use an approximative distribution $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$ Instead of the (unknown) real posterior.
  - The KL divergence measures the difference between the approximative distribution $q(\mathbf{Z})$ and the real posterior $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$.
  - In every EM iteration, we try to make this difference smaller.

Recap: Analysis of EM

- Visualization in the space of parameters



- The EM algorithm alternately
  - Computes a lower bound on the log-likelihood for the current parameters values
  - And then maximizes this bound to obtain the new parameter values.

Topics of This Lecture

- Recap: General EM
- Bayesian Estimation Revisited
  - Conjugate priors
  - Probability distributions
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
- Approximate Inference for Bayesian Mixture Models
  - Gibbs Sampler

Motivation

- Recall: Bayesian estimation

$$p(x|X) = \int p(x|\theta)\frac{p(X|\theta)p(\theta)}{\int p(X|\theta')p(\theta')\mathrm{d}\theta'}\mathrm{d}\theta$$

  - So far, we have only done this for Gaussian distributions, where the integrals could be solved analytically.
  - Now, let's also examine other distributions…

Image created with Tagxedo.com

Conjugate Priors

- Problem: How to evaluate the integrals?
  - We will see that if likelihood and prior have the same functional form $c \cdot f(x)$, then the analysis will be greatly simplified and the integrals will be solvable in closed form.

$$\begin{aligned} p(X|\theta)p(\theta) &= \prod_{x_n} c_1 f(x_n,\theta) c_2 f(\theta,\alpha) \\ &= \prod_{x_n} c f(x_n,\theta,\alpha) \end{aligned}$$

  - Such an algebraically convenient choice is called a conjugate prior. Whenever possible, we should use it.
  - To do this, we need to know for each probability distribution what is its conjugate prior. ⇒ *Topic of this lecture.*
- What to do when we cannot use the conjugate prior?
  ⇒ *Use approximate inference methods.*

The Multinomial Distribution

- Multinomial Distribution
  - Joint distribution over $m_1,\ldots,m_K$ conditioned on μ and N

$$\mathrm{Mult}(m_1,m_2,\ldots,m_K|\boldsymbol{\mu},N) = \binom{N}{m_1 m_2 \ldots m_K}\prod_{k=1}^{K}\mu_k^{m_k}$$

    with the normalization coefficient

$$\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1!m_2!\ldots m_K!}$$

  - Properties

$$\begin{aligned} \mathbb{E}[m_k] &= N\mu_k \\ \mathrm{var}[m_k] &= N\mu_k(1-\mu_k) \\ \mathrm{cov}[m_j m_k] &= -N\mu_j\mu_k \end{aligned}$$

Slide adapted from C. Bishop

Bayesian Multinomial

- Conjugate prior for the Multinomial
  - Introduce a family of prior distributions for the parameters $\{\mu_k\}$ of the Multinomial.
  - The conjugate prior is given by

$$p(\mu|\alpha) \propto \prod_{k=1}^{K}\mu_k^{\alpha_k-1}$$

    with the constraints

$$\forall k: 0 \le \mu_k \le 1 \quad \text{and} \quad \sum_{k=1}^{K}\mu_k = 1$$

The Dirichlet Distribution

- Dirichlet Distribution
  - Multivariate generalization of the Beta distribution

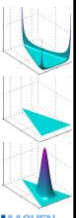
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1} \quad \text{with} \quad \alpha_0 = \sum_{k=1}^{K}\alpha_k$$

- Properties
  - Conjugate prior for the Multinomial.
  - The Dirichlet distribution over K variables is confined to a K-1 dimensional simplex.
  - Expectations:

$$\mathbb{E}[\mu_k] = \frac{\alpha_k}{\alpha_0}$$

$$\mathrm{var}[\mu_k] = \frac{\alpha_k(\alpha_0-\alpha_k)}{\alpha_0^2(\alpha_0+1)}$$

$$\mathrm{cov}[\mu_j\mu_k] = -\frac{\alpha_j\alpha_k}{\alpha_0^2(\alpha_0+1)}$$



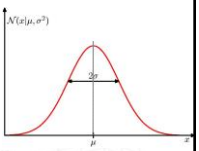
Slide adapted from C. Bishop

Image source: C. Bishop, 2006

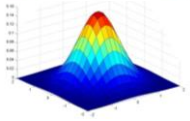
Recap: The Gaussian Distribution

- One-dimensional case
  - Mean μ
  - Variance σ^2

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



- Multi-dimensional case
  - Mean μ
  - Covariance $\boldsymbol{\Sigma}$



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Image source: C.M. Bishop, 2006

Bayesian Inference for the Gaussian

- Univariate conjugate priors
  - σ^2 known, μ unknown: $p(\mu)$ Gaussian

$$p(\mathbf{X}|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

  - μ is known, λ unknown: $p(\lambda)$ Gamma

$$p(\mathbf{X}|\lambda) \propto \lambda^{N/2}\exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

  - both μ and λ unknown: $p(\mu,\lambda)$ Gaussian-Gamma

$$p(\mathbf{X}|\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.$$

Slide adapted from C. Bishop

Bayesian Inference for the Gaussian

- Multivariate conjugate priors
  - μ unknown, Λ known: $p(\mu)$ Gaussian.
  - Λ unknown, μ known: $p(\Lambda)$ Wishart,

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu) = B|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right).$$

  - Λ and μ unknown: $p(\mu,\Lambda)$ Gaussian-Wishart,

$$p(\mu,\boldsymbol{\Lambda}|\mu_0,\beta,\mathbf{W},\nu) = \mathcal{N}(\mu|\mu_0,(\beta\boldsymbol{\Lambda})^{-1})\,\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu)$$

Slide adapted from C. Bishop

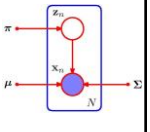
Topics of This Lecture

- Recap: General EM
- Bayesian Estimation Revisited
  - Conjugate priors
  - Probability distributions
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
- Approximate Inference for Bayesian Mixture Models
  - Gibbs Sampler

Towards a Full Bayesian Treatment…

- Mixture models
  - We have discussed mixture distributions with K components

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$$



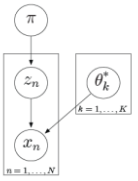
  - So far, we have derived the ML estimates ⇒ EM
  - Introduced a prior $p(\theta)$ over parameters ⇒ MAP-EM
  - One question remains open: how to set K ?
    ⇒ Let's also set a prior on the number of components…

Bayesian Mixture Models

- Let's be Bayesian about mixture models
  - Place priors over our parameters
  - Again, introduce variable $\mathbf{z}_n$ as indicator which component data point $\mathbf{x}_n$ belongs to.

$$\mathbf{z}_n|\boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi})$$
$$\mathbf{x}_n|\mathbf{z}_n = k, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

  - This is similar to the graphical model we've used before, but now the π and $\theta_k = (\mu_k, \boldsymbol{\Sigma}_k)$ are also treated as random variables.
  - *What would be suitable priors for them?*



Bayesian Mixture Models

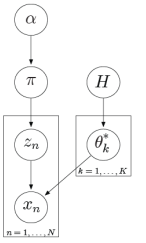
- Let's be Bayesian about mixture models
  - Place priors over our parameters
  - Again, introduce variable $\mathbf{z}_n$ as indicator which component data point $\mathbf{x}_n$ belongs to.

$$\mathbf{z}_n|\boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi})$$
$$\mathbf{x}_n|\mathbf{z}_n = k, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

  - Introduce conjugate priors over parameters

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$$
$$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim H = \mathcal{NIW}(0, s, d, \phi)$$

"Normal – Inverse Wishart"



Bayesian Mixture Models

- Full Bayesian Treatment
  - Given a dataset, we are interested in the cluster assignments

$$p(\mathbf{Z}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{\sum_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}$$

  where the likelihood is obtained by marginalizing over the parameters θ

$$p(\mathbf{X}|\mathbf{Z}) = \int p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$
$$= \int \prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_n|z_{nk}, \boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|H)\mathrm{d}\boldsymbol{\theta}$$

- The posterior over assignments is intractable!
  - Denominator requires summing over all possible partitions of the data into K groups!

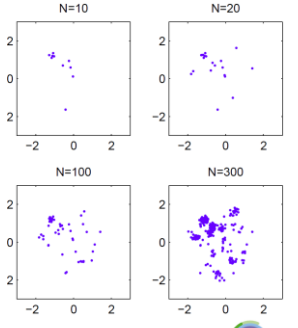
⇒ Need efficient approximate inference methods to solve this...

Bayesian Mixture Models

- Let's examine this model more closely
  - Role of Dirichlet priors?
  - How can we perform efficient inference?
  - What happens when K goes to infinity?

- This will lead us to an interesting class of models…
  - Dirichlet Processes
  - Possible to express infinite mixture distributions with their help
  - Clustering that automatically adapts the number of clusters to the data and *dynamically creates new clusters on-the-fly*.

Sneak Preview: Dirichlet Process MoG



Samples drawn from DP mixture

⇒ More structure appears as more points are drawn

Recap: The Dirichlet Distribution

- Dirichlet Distribution
  - Conjugate prior for the Categorical and the Multinomial distrib.

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \quad \text{with} \quad \alpha_0 = \sum_{k=1}^{K} \alpha_k$$

  - Symmetric version (with concentration parameter α)

$$\text{Dir}(\boldsymbol{\mu}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \mu_k^{\alpha/K - 1}$$

  - Properties (symmetric version)

$$\mathbb{E}[\mu_k] = \frac{\alpha_k}{\alpha_0} = \frac{1}{K}$$
$$\text{var}[\mu_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} = \frac{K-1}{K^2(\alpha+1)}$$
$$\text{cov}[\mu_j \mu_k] = -\frac{\alpha_j \alpha_k}{\alpha_0^2(\alpha_0 + 1)} = -\frac{1}{K^2(\alpha+1)}$$

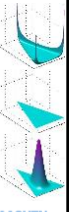
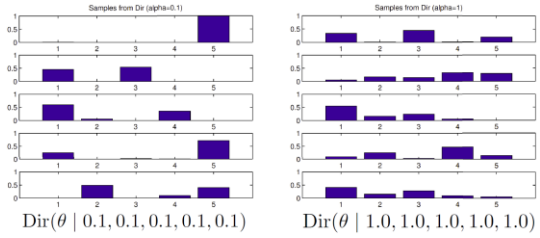


Image source: C. Bishop, 2006

Dirichlet Samples



$\mathrm{Dir}(\theta \mid 0.1, 0.1, 0.1, 0.1, 0.1)$ $\mathrm{Dir}(\theta \mid 1.0, 1.0, 1.0, 1.0, 1.0)$

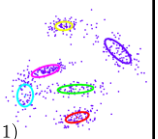
- Effect of concentration parameter α
  - Controls sparsity of the resulting samples

Slide credit: Erik Sudderth
Image source: Erik Sudderth

Mixture Model with Dirichlet Priors

- Finite mixture of K components

$$p(\mathbf{x}_n|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\theta_k)$$
$$= \sum_{k=1}^{K} p(z_{nk}=1|\pi_k) p(\mathbf{x}_n|\theta_k, z_{nk}=1)$$



  - The distribution of latent variables $\mathbf{z}_n$ given π is multinomial

$$p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{N_k}, \quad N_k \stackrel{\text{def}}{=} \sum_{n=1}^{N} z_{nk}$$

  - Assume mixing proportions have a given symmetric conjugate Dirichlet

$$p(\boldsymbol{\pi}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{\alpha/K-1}$$

Slide adapted from Zoubin Ghahramani
Image source: Zoubin Ghahramani

Mixture Model with Dirichlet Priors

- Integrating out the mixing proportions π:

$$p(\mathbf{z}|\alpha) = \int p(\mathbf{z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\alpha) \mathrm{d}\boldsymbol{\pi}$$
$$= \int \prod_{k=1}^{K} \pi_k^{N_k} \cdot \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{\alpha/K-1} \mathrm{d}\boldsymbol{\pi}$$
$$= \int \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{N_k+\alpha/K-1} \mathrm{d}\boldsymbol{\pi}$$

  - This is again a Dirichlet distribution (reason for conjugate priors)

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \frac{\prod_{k=1}^{K} \Gamma(N_k+\alpha/K)}{\Gamma(N+\alpha)} \int \frac{\Gamma(N+\alpha)}{\prod_{k=1}^{K} \Gamma(N_k+\alpha/K)} \prod_{k=1}^{K} \pi_k^{N_k+\alpha/K-1} \mathrm{d}\boldsymbol{\pi}$$

Completed Dirichlet form → integrates to 1

Mixture Models with Dirichlet Priors

- Integrating out the mixing proportions π (cont'd)

$$p(\mathbf{z}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \frac{\prod_{k=1}^{K} \Gamma(N_k+\alpha/K)}{\Gamma(N+\alpha)}$$
$$= \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k+\alpha/K)}{\Gamma(\alpha/K)}$$

- Conditional probabilities
  - Let's examine the conditional of $\mathbf{z}_n$ given all other variables

$$p(z_{nk}=1|\mathbf{z}_{-n}, \alpha) = \frac{p(z_{nk}=1, \mathbf{z}_{-n}|\alpha)}{p(\mathbf{z}_{-n}|\alpha)}$$

where $\mathbf{z}_{-n}$ denotes all indizes except n.

Slide adapted from Zoubin Ghahramani

Mixture Models with Dirichlet Priors

- Conditional probabilities

$$p(\mathbf{z}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k+\alpha/K)}{\Gamma(\alpha/K)}$$

$$p(z_{nk}=1|\mathbf{z}_{-n}, \alpha) = \frac{p(z_{nk}=1, \mathbf{z}_{-n}|\alpha)}{p(\mathbf{z}_{-n}|\alpha)}$$
$$= \frac{\cancel{\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}} \frac{\Gamma(N_k+\alpha/K)}{\cancel{\Gamma(\alpha/K)}} \cancel{\prod_{j=1, j\neq k}^{K} \frac{\Gamma(N_j+\alpha/K)}{\Gamma(\alpha/K)}}}{\cancel{\frac{\Gamma(\alpha)}{\Gamma(N_{-n}+\alpha)}} \frac{\Gamma(N_{-n,k}+\alpha/K)}{\cancel{\Gamma(\alpha/K)}} \cancel{\prod_{j=1, j\neq k}^{K} \frac{\Gamma(N_j+\alpha/K)}{\Gamma(\alpha/K)}}}$$
$$= \frac{\Gamma(N_{-n}+\alpha)}{\Gamma(N+\alpha)} \frac{\Gamma(N_k+\alpha/K)}{\Gamma(N_{-n,k}+\alpha/K)}$$

Mixture Models with Dirichlet Priors

- Conditional probabilities

$$\Gamma(n+1) = n\Gamma(n)$$

$$p(z_{nk}=1|\mathbf{z}_{-n}, \alpha) = \frac{p(z_{nk}=1, \mathbf{z}_{-n}|\alpha)}{p(\mathbf{z}_{-n}|\alpha)}$$
$$= \frac{\cancel{\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}} \frac{\Gamma(N_k+\alpha/K)}{\cancel{\Gamma(\alpha/K)}} \cancel{\prod_{j=1, j\neq k}^{K} \frac{\Gamma(N_j+\alpha/K)}{\Gamma(\alpha/K)}}}{\cancel{\frac{\Gamma(\alpha)}{\Gamma(N_{-n}+\alpha)}} \frac{\Gamma(N_{-n,k}+\alpha/K)}{\cancel{\Gamma(\alpha/K)}} \cancel{\prod_{j=1, j\neq k}^{K} \frac{\Gamma(N_j+\alpha/K)}{\Gamma(\alpha/K)}}}$$
$$= \frac{\Gamma(N_{-n}+\alpha)}{\Gamma(N+\alpha)} \frac{\Gamma(N_k+\alpha/K)}{\Gamma(N_{-n,k}+\alpha/K)}$$
$$= \frac{1}{N-1+\alpha} \frac{N_{-n,k}+\alpha/K}{1}$$
$$= \frac{N_{-n,k}+\alpha/K}{N-1+\alpha}$$

Finite Dirichlet Mixture Models

- Conditional probabilities: Finite K

$$p(z_{nk}=1|\mathbf{z}_{-n},\alpha) = \frac{N_{-n,k}+\alpha/K}{N-1+\alpha}, \qquad N_{-n,k} \stackrel{\text{def}}{=} \sum_{i=1,i\neq n}^{N} z_{ik}$$

- This is a very interesting result. *Why?*
  - We directly get a numerical probability, no distribution.
  - The probability of joining a cluster mainly depends on the number of existing entries in a cluster.
    ⇒ The more populous a class is, the more likely it is to be joined!
  - In addition, we have a base probability of also joining as-yet empty clusters.
  - This result can be directly used in Gibbs Sampling…
    (*see later derivation*)

Infinite Dirichlet Mixture Models

- Conditional probabilities: Finite K

$$p(z_{nk}=1|\mathbf{z}_{-n},\alpha) = \frac{N_{-n,k}+\alpha/K}{N-1+\alpha}, \qquad N_{-n,k} \stackrel{\text{def}}{=} \sum_{i=1,i\neq n}^{N} z_{ik}$$

- Conditional probabilities: Infinite K
  - Taking the limit as $K \to \infty$ yields the conditionals

$$p(z_{nk}=1|\mathbf{z}_{-n},\alpha) = \begin{cases} \frac{N_{-n,k}}{N-1+\alpha} & \text{if } k \text{ represented} \\ \frac{\alpha}{N-1+\alpha} & \text{if all } k \text{ not represented} \end{cases}$$

  - Left-over mass α ⇒ countably infinite number of indicator settings

Discussion

- Infinite Mixture Models
  - What we have just seen is a first example of a Dirichlet Process.
  - DPs allow us to work with models that have an infinite number of components.
  - This will raise a number of issues
    - How to represent infinitely many parameters?
    - How to deal with permutations of the class labels?
    - How to control the effective size of the model?
    - How to perform efficient inference?

  ⇒ *More background needed here!*
  - DPs are a very interesting class of models, but would take us too far here.
  - If you're interested in learning more about them, take a look at the Advanced ML slides from Winter 2012.

Topics of This Lecture

- Recap: General EM
- Bayesian Estimation Revisited
  - Conjugate priors
  - Probability distributions
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
- Approximate Inference for Bayesian Mixture Models
  - Gibbs Sampler

Gibbs Sampling for Finite Mixtures

- We need approximate inference here
  - Gibbs Sampling: Conditionals are simple to compute

$$p(\mathbf{z}_n = k|\text{others}) \propto \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\boldsymbol{\pi} \mid \mathbf{z} \sim \text{Dir}(N_1+\alpha/K, \ldots, N_K+\alpha/K)$$

$$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\text{others} \sim \mathcal{NIW}(v', s', d', \phi')$$

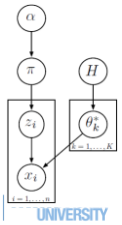
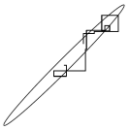


Image source: Yee Whye Teh

Recap: Gibbs Sampling

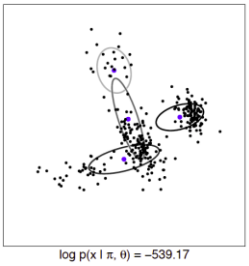
- Approach
  - MCMC-algorithm that is simple and widely applicable.
  - May be seen as a special case of Metropolis-Hastings.
- Idea
  - Sample variable-wise: replace $\mathbf{z}_i$ by a value drawn from the distribution $p(z_i|\mathbf{z}_{\setminus i})$.
    - This means we update one coordinate at a time.
  - Repeat procedure either by cycling through all variables or by choosing the next variable.
- Properties
  - The algorithm always accepts!
  - Completely parameter free.
  - Can also be applied to subsets of variables.



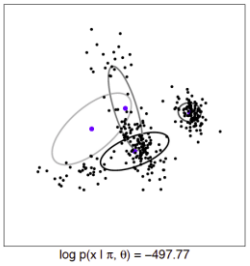
Gibbs Sampling for Finite Mixtures

- Standard finite mixture sampler
  - Given mixture weights $\pi^{(t-1)}$ and cluster parameters $\left\{\boldsymbol{\theta}_k^{(t-1)}\right\}_{k=1}^K$ from the previous iteration, sample new parameters as follows
    1. Independently assign each point $\mathbf{x}_n$ to one of the K clusters by sampling the variables $\mathbf{z}_n$ from the multinomial distributions
    $$\mathbf{z}_n^{(t)} \sim \frac{1}{Z_n}\sum_{k=1}^{K} z_{nk}^{(t-1)}\pi_k^{(t-1)}p(\mathbf{x}_n|\theta_k^{(t-1)}) \qquad Z_n = \sum_{k=1}^{K}\pi_k^{(t-1)}p(\mathbf{x}_n|\theta_k^{(t-1)})$$
    2. Sample new mixture weights from the Dirichlet distribution
    $$\boldsymbol{\pi}^{(t)} \sim \mathrm{Dir}(N_1+\alpha/K,\ldots,N_K+\alpha/K) \qquad N_k = \sum_{n=1}^{N} z_{nk}^{(t)}$$
    3. For each of the K clusters, independently sample new parameters from the conditional of the assigned observations
    $$\boldsymbol{\theta}_k^{(t)} \sim p(\boldsymbol{\theta}_k|\{\mathbf{x}_n|z_{nk}=1\}, H)$$

Standard Sampler: 2 Iterations

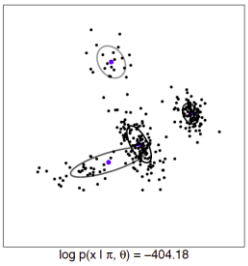


log p(x | π, θ) = –539.17

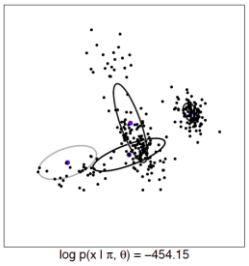


log p(x | π, θ) = –497.77

Standard Sampler: 10 Iterations

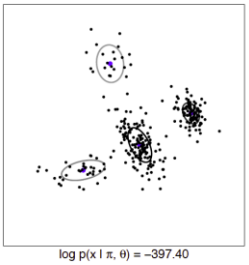


log p(x | π, θ) = –404.18

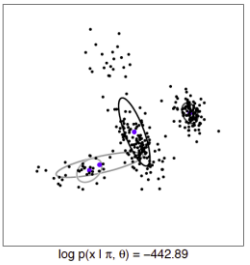


log p(x | π, θ) = –454.15

Standard Sampler: 10 Iterations



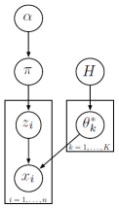
log p(x | π, θ) = –397.40



log p(x | π, θ) = –442.89

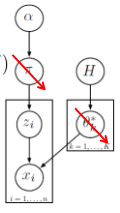
Gibbs Sampling for Finite Mixtures

- We need approximate inference here
  - Gibbs Sampling: Conditionals are simple to compute
  $$p(\mathbf{z}_n = k|\text{others}) \propto \sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$
  $$\boldsymbol{\pi} \mid \mathbf{z} \sim \mathrm{Dir}(N_1+\alpha/K,\ldots,N_K+\alpha/K)$$
  $$\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k|\text{others} \sim \mathcal{NIW}(v',s',d',\phi')$$
- However, this will be rather inefficient…
  - In each iteration, algorithm can only change the assignment for individual data points.
  - There are often groups of data points that are associated with high probability to the same component. ⇒ Unlikely that group is moved.
  - Better performance by collapsed Gibbs sampling which integrates out the parameters π, μ, $\boldsymbol{\Sigma}$.



Collapsed Finite Bayesian Mixture

- More efficient algorithm
  - Conjugate priors allow analytic integration of some parameters
  - Resulting sampler operates on reduced space of cluster assignments (implicitly considers all possible cluster shapes)
- Procedure
  - The model implies the factorization
  $$p(\mathbf{z}_n|\mathbf{z}_{-n},\mathbf{x},\alpha,H) \propto p(\mathbf{z}_n|\mathbf{z}_{-n},\alpha)p(\mathbf{x}_n|\mathbf{z},\mathbf{x}_{-n},H)$$
  - Derive
  $$p(\mathbf{z}|\alpha) = \int p(\mathbf{z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)\mathrm{d}\boldsymbol{\pi}$$



Recap: Mixture Models with Dirichlet Priors

- Integrating out the mixing proportions π

$$p(\mathbf{z}|\alpha) = \int p(\mathbf{z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)\mathrm{d}\boldsymbol{\pi}$$
$$= \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}\prod_{k=1}^{K}\frac{\Gamma(N_k+\alpha/K)}{\Gamma(\alpha/K)}$$

- Conditional probabilities
  - Examine the conditional of $\mathbf{z}_n$ given all other variables $\mathbf{z}_{-n}$

$$p(z_{nk}=1|\mathbf{z}_{-n},\alpha) = \frac{p(z_{nk}=1,\mathbf{z}_{-n}|\alpha)}{p(\mathbf{z}_{-n}|\alpha)}$$
$$= \frac{N_{-n,k}+\alpha/K}{N-1+\alpha} \qquad N_{-n,k} \stackrel{\text{def}}{=} \sum_{i=1,i\neq n}^{N} z_{ik}$$

⇒ The more populous a class is, the more likely it is to be joined!

Slide adapted from Zoubin Gharamani

Collapsed Finite Bayesian Mixture

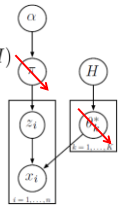
- More efficient algorithm
  - Conjugate priors allow analytic integration of some parameters
  - Resulting sampler operates on reduced space of cluster assignments (implicitly considers all possible cluster shapes)
- Procedure
  - The model implies the factorization

$$p(\mathbf{z}_n|\mathbf{z}_{-n},\mathbf{x},\alpha,H) \propto p(\mathbf{z}_n|\mathbf{z}_{-n},\alpha)p(\mathbf{x}_n|\mathbf{z},\mathbf{x}_{-n},H)$$

  - Derive

$$p(\mathbf{z}|\alpha) = \int p(\mathbf{z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)\mathrm{d}\boldsymbol{\pi}$$
$$p(\mathbf{x}_n|\mathbf{z}_n,H) = \int \sum_{k=1}^{K} z_{nk}p(\mathbf{x}_n|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|H)\mathrm{d}\boldsymbol{\theta}$$

⇒ Conjugate prior, Normal - Inverse Wishart



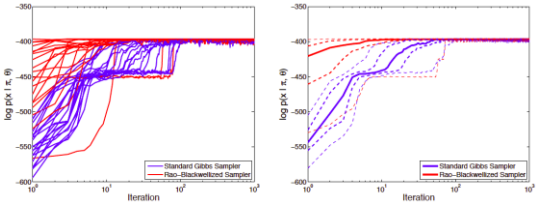
Slide adapted from Erik Sudderth
Image source: Yee Whye Teh

Collapsed (Rao-Blackwellized) Finite Mixture Sampler

- Algorithm
  1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1,\ldots,N\}$.
  2. Set $\mathbf{z} = \mathbf{z}^{(t-1)}$. For each $i \in \{\tau(1),\ldots,\tau(N)\}$, sequentially resample $\mathbf{z}_i$ as follows
     a) For each of the K clusters, determine the predictive likelihood (this can be computed from cached sufficient statistics)
     $$p_k(\mathbf{x}_n|\mathbf{z}_{-n},H) = p(\mathbf{x}_n|\{\mathbf{x}_m|z_{mk}=1,m\neq n\},H)$$
     b) Sample a new assignment $\mathbf{z}_n$ from the multinomial distribution
     $$\mathbf{z}_n \sim \sum_{k=1}^{K}\frac{z_{nk}(N_{-n,k}+\alpha/K)p_k(\mathbf{x}_n|\mathbf{z}_{-n},H)}{\sum_{j=1}^{K}(N_{-n,j}+\alpha/K)p_j(\mathbf{x}_n|\mathbf{z}_{-n},H)}$$
     c) Update cached sufficient statistics to reflect assignment z_{nk}.
  3. Set $\mathbf{z}^{(t)} = \mathbf{z}$. Optionally, mixture parameters may be sampled via steps 2-3 of the standard finite mixture sampler.

Slide adapted from Erik Sudderth

Standard vs. Collapsed Samplers



⇒ Collapsed sampler converges much more quickly.
- Theorem (Rao-Blackwell)

"Analytical marginalization of some variables from a joint distribution always reduces the variance of later estimates."

Slide credit: Erik Sudderth
Image source: Erik Sudderth

Discussion

- Collapsed Gibbs sampling
  - Integrates out the parameters π, μ, $\boldsymbol{\Sigma}$.

$$p(z_{nk}=1|\text{others}) \propto \frac{(N_{-n,k}+\alpha/K)}{N-1+\alpha}p_k(\mathbf{x}_n|\mathbf{z}_{-n},H)$$

- Properties
  - Can change all assignments in each iteration.
    ⇒ Able to move entire groups between clusters.
    ⇒ Faster convergence, less likely to get stuck.

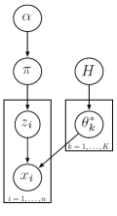


Image source: Yee Whye Teh

References and Further Reading

- Unfortunately, there are currently no good introductory textbooks on the Dirichlet Process. We therefore recommend a number of tutorial papers on their different aspects.
  - One of the best available general introductions
    - E.B. Sudderth, "Graphical Models for Visual Object Recognition and Tracking", PhD thesis, Chapter 2, Section 2.5, 2006.
  - A gentle introductory tutorial (recommended 1st read)
    - S.J. Gershman, D.M. Blei, „A Tutorial on Bayesian Nonparametric Methods", In Journal of Mathematical Psychology, Vol. 56, 2012.
  - Good overview of MCMC methods for DPMMs
    - R. Neal, Markov Chain Sampling Methods for Dirichlet Process Mixture Models, Journal of Computational and Graphical Statistics, Vol. 9(2), p. 249-265, 2000.